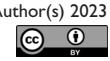
Article

LANGUAGE TESTING

Language Testing
1–39


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02655322231155561
journals.sagepub.com/home/ltj


Operationalizing the reading-into-writing construct in analytic rating scales: Effects of different approaches on rating

Santi B. Lestari
Lancaster University, UK

Tineke Brunfaut
Lancaster University, UK

Abstract

Assessing integrated reading-into-writing task performances is known to be challenging, and analytic rating scales have been found to better facilitate the scoring of these performances than other common types of rating scales. However, little is known about how specific operationalizations of the reading-into-writing construct in analytic rating scales may affect rating quality, and by extension score inferences and uses. Using two different analytic rating scales as proxies for two approaches to reading-into-writing construct operationalization, this study investigated the extent to which these approaches affect rating reliability and consistency. Twenty raters rated a set of reading-into-writing performances twice, each time using a different analytic rating scale, and completed post-rating questionnaires. The findings resulting from our convergent explanatory mixed-method research design show that both analytic rating scales functioned well, further supporting the use of analytic rating scales for scoring reading-into-writing. Raters reported that either type of analytic rating scale prompted them to attend to the reading-related aspects of reading-into-writing, although rating these aspects remained more challenging than judging writing-related aspects. The two scales differed, however, in the extent to which they led raters to uniform interpretations of performance difficulty levels. This study has implications for reading-into-writing scale design and rater training.

Indonesian

Operasionalisasi konstruk membaca-untuk-menulis ke dalam rubrik penilaian analitis: Efek dari pendekatan yang berbeda pada penilaian

Corresponding author:
Santi B. Lestari, Department of Linguistics and English Language, Lancaster University, Lancaster LA1 4YL, UK.
Email: s.lestari@lancaster.ac.uk

Menilai performa siswa dalam tes membaca-untuk-menulis sangatlah menantang, dan penelitian telah menunjukkan bahwa rubrik penilaian analitis terbukti lebih baik untuk penilaian tersebut daripada jenis rubrik penilaian yang lain. Akan tetapi, tidak banyak diketahui apakah perbedaan cara operasionalisasi konstruk membaca-untuk-menulis ke dalam rubrik penilaian analitis akan memengaruhi kualitas penilaian yang kemudian dapat berimbas pada kesimpulan yang diperoleh berdasarkan nilai dan kegunaan nilai itu sendiri dalam pengambilan keputusan. Dengan menggunakan dua rubrik penilaian analitis yang berbeda sebagai proksi dari dua pendekatan operasionalisasi konstruk membaca-untuk-menulis ke dalam rubrik penilaian analitis, studi ini meneliti seberapa jauh perbedaan pendekatan tersebut dapat memengaruhi reliabilitas dan konsistensi penilaian. Dua puluh penilai berpartisipasti dalam penelitian ini dan mereka diminta untuk menilai satu set performa membaca-untuk-menulis dua kali, setiap kali menggunakan rubrik penilaian yang berbeda dan mengisi kuesioner setelahnya. Hasil penelitian yang menggunakan desain *convergent explanatory mixed-methods* ini menunjukkan bahwa kedua rubrik penilaian analitis berfungsi dengan baik, mendukung hasil penelitian sebelumnya tentang penggunaan rubrik penilaian analitis untuk menilai performa membaca-untuk-menulis. Peserta penelitian mengatakan bahwa kedua rubrik penilaian tersebut mendorong mereka untuk memperhatikan aspek kemahiran membaca dalam membaca-untuk-menulis walaupun menilai aspek tersebut tetap lebih menantang dan rumit dibandingkan dengan menilai aspek kemahiran menulis. Kedua rubrik penilaian memiliki perbedaan dalam hal mendorong penilai untuk memiliki kesamaan pemahaman mengenai tingkat kesulitan performa. Penelitian ini memiliki implikasi pada desain soal membaca-untuk-menulis dan pelatihan menilai bagi para penilai performa ini.

Keywords
Analytic rating, integrated skills testing, many-facet Rasch measurement (MFRM), raters, rating scales, reading-into-writing

Introduction

Reading-into-writing tasks, a type of integrated task in which test-takers are presented with reading texts and required to integrate information from these texts into their writing, are increasingly adopted to either replace or supplement independent, writing-only tasks in second language assessments. Research has shown that reading-into-writing elicits, for example, composing processes similar to those employed when writing in academic contexts (Chan, 2013; Plakans, 2008, 2009) and written products which are more similar to disciplinary writing than elicited by writing-only tasks (Staples et al., 2018). Reading-into-writing tasks' authenticity is largely due to the fact that, to succeed in these tasks, test-takers are required to demonstrate their understanding of source material and their ability to appropriately and effectively use this source material in their writing (Cumming, 2013; Ohta et al., 2018). The centrality of reading comprehension and source use in the successful completion of reading-into-writing tasks thus makes evaluation of these aspects in the scoring of performances imperative (Plakans & Gebril, 2015); otherwise, rating underrepresents the construct, hence potentially raising concerns over the validity of score interpretations and uses.

The scoring of reading-into-writing performances has received growing research attention (e.g., Gebril & Plakans, 2014; Ohta et al., 2018; Shin & Ewert, 2015; J. Wang et al., 2017), and it is well established that scoring reading-into-writing performances is

highly complex, as raters have to attend to both test-takers' own written production and their uses of the source material. To facilitate the rating of reading-into-writing performances in a reliable and valid manner, different reading-into-writing rating scales have been developed and researched. Analytic rating scales have been found to result in more reliable scores than holistic ones for rating reading-into-writing performances (Ohta et al., 2018), and to increase raters' confidence in scoring such performances (Chan et al., 2015). One possible reason is that analytic rating scales for integrated testing often contain one or more distinct criteria that guide raters to make more consistent judgements on source use–related aspects specifically. However, existing analytic rating scales for reading-into-writing are not uniform in how they operationalize the construct. Some scales assign one or more specific, separate criteria to represent the reading and source use aspects *exclusively*. Other scales interweave the reading and source use aspects into criteria also capturing writing and language aspects; thus, the reading and source use aspects are spread across several, *multifaceted* rating criteria in those scales. Little is known regarding the impact on raters' rating of these different approaches to construct operationalization, which in turn may affect the validity of reading-into-writing score inferences and uses. This paper reports on a study investigating two different approaches to reading-into-writing construct operationalization in analytic rating scales and the two approaches' potential impact on rating reliability and consistency.

Literature review

The reading-into-writing construct and implications for rating

Previous research investigating the construct of reading-into-writing found that the presence of and requirement to integrate source texts in reading-into-writing tasks elicit cognitive processes and writing performance distinct from those elicited by independent writing-only tasks. In terms of cognitive processes, reading-into-writing tasks are known to uniquely elicit discourse synthesis processes (Spivey, 1997), which comprise, *organizing* ideas both while reading the source texts and composing the written response, *selecting* relevant ideas from the source texts, and *connecting* ideas across the source texts and with their own ideas (Chan, 2018; Plakans, 2008, 2009; P. Wang, 2018). Given these unique cognitive processes, reading-into-writing scores typically correlate only weakly with reading comprehension scores and independent writing scores (Delaney, 2008). Studies comparing the discourse features of reading-into-writing versus writing-only performances generally found that the former are more lexically sophisticated (Cumming et al., 2005; Gebril & Plakans, 2016) and syntactically complex (Cumming et al., 2005). However, especially regarding lexical sophistication, this might partly result from the integration of source information (Gebril & Plakans, 2016; Yu, 2013). Plakans and Gebril (2012), for example, observed that source texts assisted test-takers in idea generation as well as with language support, providing them with vocabulary, spelling, and organization models.

The need for test-takers to comprehend information from source texts and later integrate the information into their writing has been lauded as contributing to the authenticity of the reading-into-writing task type (Cumming, 2013) and, therefore, must be included in the rating scales to ensure scoring validity and valid interpretation

of scores. However, this very feature of integration poses challenges in the scoring of reading-into-writing performances (Chan et al., 2015; Cumming, 2013). First, the measurement of writing ability is confounded with that of reading ability, leading to what is often termed *task dependency* (Cumming, 2014): difficulty in source text comprehension likely impedes successful reading-into-writing task completion. It is consequently often argued that integrated tasks require certain thresholds of proficiency to allow meaningful performance (Cumming, 2014), raising questions about the usefulness of such tasks for lower proficiency levels. Second, raters, especially human raters, are required not only to attend to the language produced by test-takers (as when scoring writing-only performances) but also to distinguish text derived from the source texts from test-takers' own text and then to evaluate whether the textual borrowing practices are appropriate and effective for the task requirements (Cumming, 2014). Research investigating the scoring processes of reading-into-writing performances has collectively shown the complexity of these processes. For example, raters in Cumming et al.'s (2002) study reported evaluating test-takers' understanding of the source texts and the appropriateness and effectiveness of source use in relation to the task, but at the same time, they also admitted requiring more guidance on how to do so. Likewise, raters in Gebril and Plakans' (2014) study reported evaluating various aspects of source use (e.g., accuracy, relevance, adequacy, effectiveness, and appropriacy) and facing difficulties distinguishing between text borrowed from the source texts and test-takers' own language production. Furthermore, J. Wang et al. (2017) found that textual borrowing was a source of rating inaccuracy, defined as the difference between scores awarded by professional raters and so-called criterion scores from expert raters.

Previous studies on reading-into-writing rating scales

Several studies have reported on the development and use of rating scales for assessing reading-into-writing that specifically include reading-related aspects. For example, Chan et al. (2015) reported on the development and validation of analytic reading-into-writing rating scales for a suite of reading-into-writing tasks at four Common European Framework of Reference for Languages (CEFR) levels. The resulting rating scales had four rating criteria, one of which, *Reading for Writing*, was specifically designed to operationalize the reading and source use aspects of the reading-into-writing construct, while the remaining three criteria focused more on writing and language aspects. In the initial round of piloting, raters reported difficulties using the new criterion, *Reading for Writing.* Nonetheless, the analytic rating scales were well received by the raters and were considered to provide more explicit guidance for rating compared to the holistic rating scales they had previously used; therefore, the analytic scales increased the raters' confidence in their marking. Quantitative analysis of the scores awarded by the raters also suggested that the analytic rating scales facilitated consistent and reliable marking, including for the *Reading for Writing* criterion.

Shin and Ewert (2015), in an investigation of what contributes to reading-into-writing performance, developed an analytic rating scale comprising five criteria with two of these specifically representing reading and source use aspects, respectively, named *Viewpoint Recognition* and *Text Engagement*. In Shin and Ewert's study, the raters were found to use the rating scale fairly consistently across the rating criteria. However, rater reliability on the reading and source use–related criteria was lower than that on the other

criteria. As Shin and Ewert did not collect qualitative data on the raters' perceptions of the scale, no explanation was offered to elucidate this quantitative finding. Nonetheless, evidence from this study further indicates the complex nature of scoring reading-into-writing, particularly with regard to the reading and source use aspects.

Holistic rating scales have also been used for scoring integrated writing. A well-known example, which has been used in a number of empirical studies, is the Test of English as a Foreign Language (TOEFL) iBT holistic rating scale for the test's reading-listening-writing task (see https://www.ets.org/content/dam/ets-org/pdfs/toefl/toefl-ibt-writing-rubrics.pdf). Ohta et al. (2018), for example, compared an adapted version of this holistic scale with an analytic one comprising five rating criteria (*Source Use, Organization, Development of Ideas, Language Use*, and *Authorial Voice*). The findings suggested that scores resulting from using the analytic rating scale were more reliable than those from the adapted TOEFL iBT holistic rating scale. Notably, the *Source Use* criterion was found to be the most reliable and consistently used criterion of the rating scale. This contradicts other similar studies which found that source use caused significant variability among raters (Gebril & Plakans, 2014), that textual borrowing was one source of rating inaccuracy (J. Wang et al., 2017), and that rater reliability in source use–related criteria was lower than that in writing and language-related criteria (Shin & Ewert, 2015). As Ohta et al.'s (2018) study was purely quantitative, it did not explore further how the analytic rating scale used in the study was able to facilitate raters to reliably score the reading and source use aspects.

Usability and variability of analytic rating scales for reading-into-writing tasks

Analytic rating scales appear to be more commonly used than other types of rating scales for reading-into-writing tasks. Findings from some of the studies reviewed above (Chan et al., 2015; Ohta et al., 2018; Shin & Ewert, 2015) have further shown the usability of analytic rating scales for guiding raters to mark reading-into-writing performances more reliably and consistently, particularly on the reading and source use aspects. However, it is also apparent that analytic rating scales for reading-into-writing tasks vary in how they operationalize the reading and source use construct in the scales. This is partly because identifying and evaluating the role of reading-related aspects of the reading-into-writing construct is rather complicated (Chan et al., 2015). Reading-related aspects can be operationalized in one or more specific rating criteria exclusively focusing on reading-related aspects, for example, a *Reading for Writing* criterion, while other criteria in the rating scale are free from any mention related to reading. An example of this can be found in the analytic rating scale developed by Chan et al. (2015; see Figure 1). Alternatively, reading-related aspects can be operationalized in multiple criteria, with some of these aspects also interwoven into criteria related to writing and linguistic resources, to account for the influence of source texts on various aspects of writing performances as identified in the literature. Such an example is the analytic rating scale developed by Plakans and Gebril (2015; see Figure 1), also used in Ohta et al. (2018).

While both types of analytic rating scales were found to function favourably with reading-into-writing tasks in these studies, questions emerged as to the extent to which these different ways of operationalization affect raters' rating. In addition, as with other analytic rating scales, it is important to investigate the extent to which the criteria within

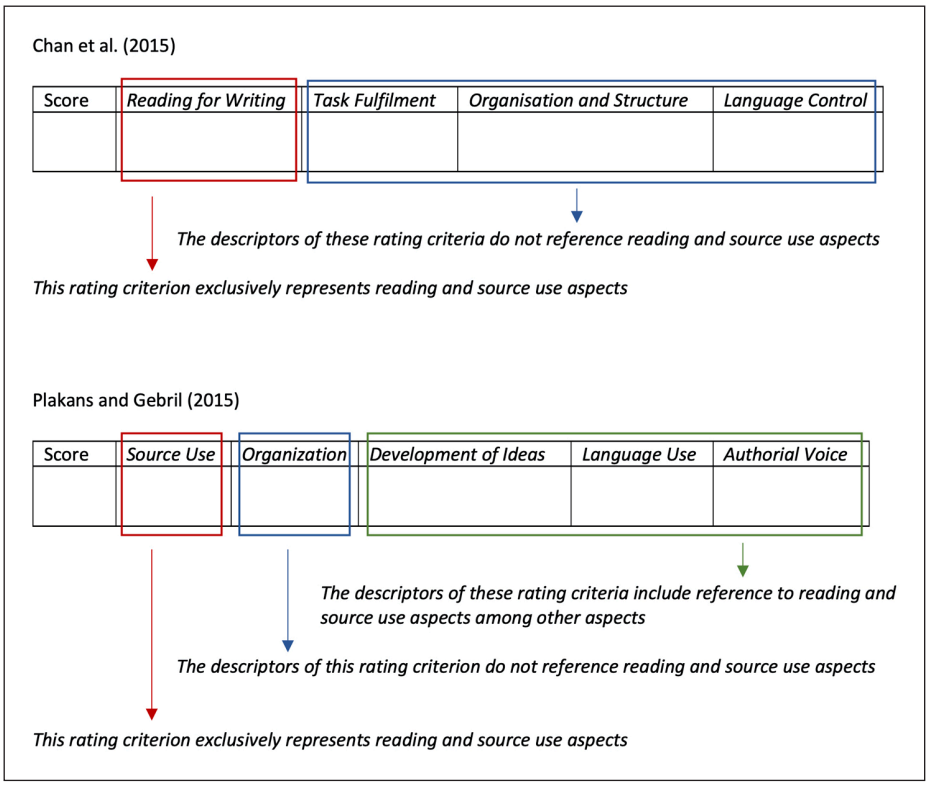


Figure 1. Reading-into-writing operationalization in analytic rating scales.

an analytic rating scale are distinguishable while still converging, to warrant their usefulness. Analytic rating scales are known to have issues with fuzziness, that is, lacking distinction across rating criteria (Knoch et al., 2021); therefore, studies examining the usability of analytic rating scales also investigated the distinguishability of criteria (e.g., Brown, 2006; Xi & Mollaun, 2006). At the same time, the different criteria within an analytic rating scale also need to converge (Knoch et al., 2020), especially when only the overall or composite score is reported, as is often the case.

Research questions

The utility of analytic rating scales for reading-into-writing assessment requires further investigation, especially given the fact that they often differ in the way they operationalize the reading-into-writing construct. Using two different analytic rating scales as proxies for two approaches to reading-into-writing construct operationalization, this study investigates the extent to which these different approaches potentially affect raters' rating reliability and consistency. More specifically, the study seeks to do so (1) quantitatively by collecting evidence of rating reliability and consistency at the level of each rating scale as a whole and at the level of each individual rating criterion within each scale, as

well as (2) qualitatively by gathering raters' perceptions of each rating scale. Therefore, four research questions were posed:

RQ1. To what extent do individual raters vary in applying two different analytic rating scales for assessing reading-into-writing?

RQ2. How do raters apply the two analytic rating scales across individual rating criteria?

RQ3. To what extent do each analytic scale's ratings provide unique but complementary information?

RQ4. What are raters' perceptions of the two analytic rating scales?

While we appreciate that rating reliability and consistency have a direct impact on the evaluation of test-takers' performance, in this paper, we focus on the raters' use of the rating scales rather than on scoring consequences.

Methods

To address the research questions, a convergent explanatory mixed-method research design was implemented, taking advantage of both quantitative and qualitative methods to arrive at a more complete understanding of the issue under investigation (Creswell & Plano Clark, 2018). The overall research design is presented in Figure 2. The study comprised mainly a quantitative method involving multiple statistical analyses of rating data to address RQ1 to RQ3, and a combination of quantitative and qualitative analyses of post-rating questionnaire responses to address RQ4 and elucidate the results of RQ1 to RQ3.

Raters

Twenty raters (24–71 years old; $M = 38$) participated in this study. Thirty percent were males, 70% were females. The majority were English first language (L1) speakers (65%), while 35% were L1 speakers of Finnish, Greek, Italian, Spanish, or Turkish. Eighty-five percent held a master's degree in fields such as (Applied) Linguistics or English Language, with most pursuing a PhD in (Applied) Linguistics at the time of data collection. Fifteen percent held a BA and/or PGCE as their highest qualification. All raters had taught English as a Foreign Language (EFL) (2–32 years; $M = 11.6$). Thirty percent had also worked as professional raters for high-stakes standardized English language tests, but none were familiar with the rating scales used in this study before participation. The raters were recruited through personal and professional contacts.

Instruments

Reading-into-writing task and performances. The task used in this study was a reading-into-writing task of the ISE II exam (CEFR B2) from Trinity College London. This task requires test-takers to write a 150- to 180-word magazine article on a popular science topic using only information presented in four reading source texts (about 500 words in

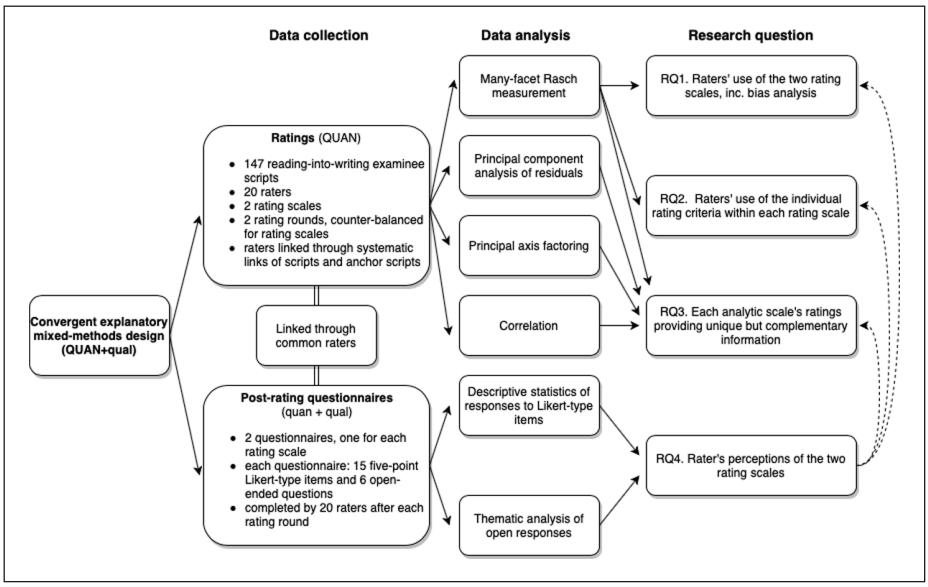


Figure 2. Research design.

total). The task aims to assess the ability to identify information from the texts relevant to the prompt; identify common themes and links across the four source texts; and paraphrase, summarize, and synthesize information from the source texts (Trinity College London, 2015). A sample task (Task 3; Trinity College London, n.d., p. 7) can be found on https://www.trinitycollege.com/resource/?id=7202. For test security reasons, the specific prompt used in this study cannot be shared here.

A total of 159 test-taker performances on the task were used in the present study: 12 scripts for rater training and 147 for rating data collection. These had been written by test-takers under live exam conditions and were sampled from approximately 500 scripts to which the researchers were given access. The sampling criteria used were score distribution based on ISE II operational raters' scores on the live test administration (to represent all four ISE II performance levels), legibility of test-takers' handwriting, and text length to ensure an adequate amount of rateable language. The sample constituted a well-balanced ratio of scripts across ISE Score Levels 1 to 3, but limited scripts at Level 4, which represents the typical ISE II test-taker population.

Rating scales. Two analytic rating scales specifically designed for rating reading-into-writing were used in this study. The two scales were deemed comparable in terms of target construct, especially in that the two are used to score reading-into-writing tasks involving *multiple* source texts, thereby targeting the important construct of discourse synthesis. At the same time, the scales differ fundamentally in their source use construct operationalization (the variable of interest in this study). The first rating scale (henceforth Rating Scale 1, Figure 3) was adapted from Trinity College London's ISE II reading-into-writing scale developed by Chan et al. (2015). The adaptations comprised (1)

Score	**Reading for writing** • Understanding of source materials • Selection of relevant content from source texts • Ability to identify common themes and links within and across the multiple texts • Adaptation of content to suit the purpose for writing • Use of paraphrasing/summarising	**Task fulfillment** • Overall achievement of communicative aim • Awareness of the writer-reader relationship (style and register) • Adequacy of topic coverage
4	• Full and accurate understanding of the essential meaning of all source materials demonstrated • A wholly appropriate and accurate selection of relevant content from the source texts • Excellent ability to identify common themes and links within and across the multiple texts and the writers' stances • An excellent adaptation of content to suit the purpose for writing • Excellent paraphrasing/summarising skills of factual ideas, opinions, argument and/or discussion demonstrated	• Excellent achievement of the communicative aim • Excellent awareness of the writer–reader relationship (ie appropriate use of standard style and register throughout the text) • All requirements (ie genre, topic, reader, purpose and number of words) of the instruction appropriately met
3	• Full and accurate understanding of the essential meaning of most source materials demonstrated • An appropriate and accurate selection of relevant content from the source texts (ie most relevant ideas are selected and most ideas selected are relevant) • Good ability to identify common themes and links within and across the multiple texts and the writers' stances • A good adaptation of content to suit the purpose for writing (eg apply the content of the source texts appropriately to offer solutions, offer some evaluation of the ideas based on the purpose for writing) • Good paraphrasing/summarising skills of factual ideas, opinions, argument and/or discussion demonstrated (with very limited lifting and few disconnected ideas)	• Good achievement of the communicative aim (ie easy to follow and convincing for reader) • Good awareness of the writer–reader relationship (ie appropriate use of standard style and register throughout the text) • Most requirements (ie, genre, topic, reader, purpose and number of words) of the instruction appropriately met
2	• Full and accurate understanding of more than half of the source materials demonstrated • An acceptable selection of relevant content from the source texts (the content selected must come from more than one text) • Acceptable ability to identify common themes and links within and across the multiple texts and the writers' stances (eg ability to discern when the same idea has been mentioned in several texts and therefore avoid repeating it) • Acceptable adaptation of content to suit the purpose for writing • Acceptable paraphrasing/summarising skills of factual ideas, opinions, argument and/or discussion demonstrated	• Acceptable achievement of the communicative aim • Some awareness of the writer–reader relationship • Most requirements (ie genre, topic, reader, purpose and number of words) of the instruction acceptably met
1	• Inaccurate and limited understanding of most source materials • Inadequate and inaccurate selection of relevant content from the source texts (ie fewer than half of the relevant ideas are selected and most of the selected ideas are irrelevant) • Poor ability to identify common themes and links within and across the multiple texts and the writers' stances (ie misunderstanding of the common themes and links is evident) • Poor adaptation of content to suit the purpose for writing (ie does not use the source texts' content to address the purpose for writing) • Poor paraphrasing/summarising skills of factual ideas, opinions, argument and/or discussion (with heavy lifting and many disconnected ideas)	• Poor achievement of the communicative aim (ie difficult to follow and unconvincing for reader) • Poor awareness of the writer–reader relationship • Most requirements (ie genre, topic, reader, purpose and number of words) of the instruction are not met

Figure 3. *(Continued)*

Score	**Organisation and structure** • Text organization, including use of paragraphing, beginnings/endings • Presentation of ideas and arguments, including clarity and coherence of their development • Consistent use of format to suit the task • Use of signposting	**Language control** • Range and accuracy of grammar • Range and accuracy of lexis • Effect of linguistic errors on understanding • Control of punctuation and spelling
4	• Effective organisation of text • Very clear presentation and logical development of most ideas and arguments, with appropriate highlighting of significant points and relevant supporting detail • Appropriate format throughout the text • Effective signposting	• Wide range of grammatical items relating to the task with good level of accuracy • Wide range of lexical items relating to the task with good level of accuracy • Any errors do not impede understanding • Excellent spelling and punctuation
3	• Good organisation of text (eg appropriately organized into clear and connected paragraphs, appropriate opening and closing) • Clear presentation and logical development of most ideas and arguments, with appropriate highlighting of significant points and relevant supporting detail • Appropriate format in most of the text • Good signposting (eg appropriate use of cohesive devices and topic sentences)	• Appropriate range of grammatical items relating to the task with good level of accuracy (with mostly non-systematic errors) • Appropriate range of lexical items relating to the task with good level of accuracy (without frequent repetition) • Errors only occasionally impede understanding • Good spelling and punctuation (may show some signs of first language influence)
2	• Acceptable organisation of text • Presentation and development of most ideas and arguments are acceptably clear and logical, with some highlighting of significant points and relevant supporting detail • Appropriate format in general • Acceptable signposting (eg some inconsistent/ faulty use of cohesive devices and topic sentences)	• Acceptable level of grammatical accuracy and appropriacy relating to the task, though range may be restricted • Acceptable level of lexical accuracy and appropriacy relating to the task, though range may be restricted • Errors sometimes impede understanding • Acceptable spelling and punctuation
1	• Very limited or poor text organisation • Most ideas and arguments lack coherence and do not progress logically • Inappropriate format throughout the text • Poor signposting (eg inappropriate or poor use of cohesive devices and topic sentences)	• Inadequate evidence of grammatical range and accuracy (may have control over the language below the level) • Inadequate evidence of lexical range and accuracy (may have control over the language below the level) • Errors frequently impede understanding • Poor spelling and punctuation throughout

Figure 3. Rating Scale 1.

Score	Source use	Organization	Development of ideas	Language use
4	• The writing shows high integration quality of source text content. • All source details included are accurately presented. • Source use is relevant and effective to address the task.	• The writing has a clear, logical and effective organizational plan. • It makes effective use of cohesive devices. • It has a solid introduction and conclusion.	• Full development of ideas. • Different types of details from the source texts are used to support the ideas and argument.	• Few language errors and language still accurately presents source ideas. • The writing includes a variety of sophisticated vocabulary and structures.
3	• The writing shows good integration quality of source text content. • Most source details included are accurately presented. • Source use is generally relevant and effective to address the task.	• The writing generally has an adequate organizational plan. • It makes good use of cohesive devices. • It has a clear introduction and conclusion.	• Development of ideas is adequate. • Details from the source texts are generally used to support the ideas and argument.	• Some language errors and they do not result in misrepresentation of source ideas. • Varied vocabulary and structures, but redundancy is sometimes an issue.
2	• The writing shows only part-integration of source text content. • Sometimes source details included are misrepresented. • Source use is somewhat relevant, but not effective to address the task.	• The organizational plan is not clear enough. • Cohesive devices are sometimes used. • There is an introduction and a conclusion.	• Development of ideas is emerging. • Few details from the source texts are used to support the ideas and argument.	• Language errors do not usually interfere with understanding meaning, but may misrepresent the source ideas. • Limited variety and common redundancy of vocabulary and structures.
1	• The writing shows inadequate integration of source text content. • Serious problems with the accuracy of source information and/or frequent verbatim use of source texts. • Source use is irrelevant and ineffective to address the task.	• The organizational plan is weak. • Few cohesive devices are used. • The argument is not easily followed.	• Little development of ideas in the writing. • Hardly any details from the source texts are used to support the ideas and argument.	• Frequent language errors interfere with understanding the writing and misrepresent source ideas. • Basic vocabulary and redundant structures.

Figure 4. Rating Scale 2.

deleting Band 0 (as no scripts in this study were at this band level) and (2) modifying the scale layout. The scale has four rating criteria: *Reading for Writing, Task Fulfilment, Organization and Structure*, and *Language Control.* The reading-related and source use construct is operationalized exclusively in one specific criterion, *Reading for Writing*. The second rating scale (henceforth Rating Scale 2, Figure 4) was adapted from a reading-into-writing rating scale developed by Plakans and Gebril (2015). The modifications comprised (1) removing one criterion—*Authorial Voice*—because the task used in this study did not require test-takers to present their personal view on the topic discussed, making this criterion irrelevant and (2) modifying some of the scale descriptors to suit the ISE II reading-into-writing task. This scale also has four rating criteria: *Source Use, Organization, Development of Ideas*, and *Language Use.* The reading-related and source use construct in this scale is operationalised in *Source Use*, but also interwoven in the descriptors under *Development of Ideas* and *Language Use*, reflecting the idea that test-takers' use of sources affects their topic development and language production.

Rater training material. Two online, self-paced rater training modules (one for each rating scale) were developed to familiarize raters with the reading-into-writing task and the relevant rating scale. Delivered via Qualtrics, each training module consisted of three parts: (A) task familiarization, (B) rating scale familiarization, and (C) rating practice. Part A required reading the four source texts of the ISE II task and answering comprehension questions. Part B comprised activities to match and order rating scale descriptors. Part C involved two rounds of rating practice: first, raters were asked to mark three scripts, write short rationales for the marks they awarded, and then compare their marks and rationales with those of an expert rater. Next, raters were assigned another three scripts for marking and do the same as in Round 1.

Post-rating questionnaires. Two post-rating questionnaires were designed to gather raters' perceptions of the usability of the rating scales, to shed light on the quantitative findings from the analyses of the rating data. Each questionnaire comprised two sections: 15 five-point Likert-type items and six open-ended questions. The Likert-type items were designed to gauge raters' perceptions of (1) the clarity of descriptors in terms of helping them mark consistently and reliably on the various criteria in the rating scales, and (2) the general practicality and usefulness of the rating scales. The items can be found in the "Results" section, Table 12. The open-ended questions of the post-rating questionnaires aimed to elicit raters' views on (1) the distinguishability of the criteria, (2) usefulness of features, and (3) challenges while using the scales:

1. Did you find the criteria adequately distinguishable? Please elaborate.
2. Are there any particular feature(s) of the rating scale that you found useful to help you mark the performances? Please explain.
3. Are there any particular feature(s) of the rating scale that you found confusing or not very helpful or difficult to operationalize? Please explain.
4. Is there anything that you feel is important but not covered in the rating scale?
5. Are there any challenges that you faced while marking the performances using the rating scale?
6. Please write here any other comments on the rating scale.

Data collection

Data collection spanned two rating rounds, one for each scale, with a 4- to 6-week interval in-between. A counter-balanced design was used to mitigate potential order effects; raters were randomly assigned to either Group A or Group B, with Group A using Rating Scale 1 in the first round and Rating Scale 2 in the second round, and Group B, the opposite. Adopting an incomplete rating design combining systematic links and anchor performances (Wind & Jones, 2019), each rater was assigned 35 scripts for marking, 7 of which were common scripts assigned to every rater (see Figure 5). In each rating round, raters had to complete the relevant rater training module, mark 35 scripts, and fill out the relevant post-rating questionnaire. Raters were given a maximum of 1 week to complete all three activities, and the majority completed each round within one or two consecutive days. This resulted in two types of data: (1) reading-into-writing scores generated from using the two rating scales and (2) post-rating questionnaire responses.

Script	1, 22, 43, 64, 85, 106, 127	2	3	4	5	…	147
Rater 1	X						
Rater 2	X	X					
Rater 3	X	X	X				
Rater 4	X	X	X	X			
Rater 5	X	X	X	X	X		
Rater 6	X		X	X	X		
Rater 7	X			X	X		X
…	X						
Rater 20	X						

Figure 5. Rating design.

Data analysis

Rating data. The rating data gathered with each rating scale were separately subjected to 3 three-facet (i.e., scripts, raters, rating scale criteria) many-facet Rasch measurement (MFRM) analyses (Linacre, 1989) using Facets software version 3.71.4 (Linacre, 2015a). First, the Rating Scale Model (Andrich, 1978) was applied to the data to examine the extent to which the rating data fit the MFRM model. MFR-RSM assumes that all the criteria within an analytic rating scale have the same threshold parameters (Eckes, 2015; McNamara et al., 2019) and is commonly used to inspect the overall functioning of analytic rating scales. The mathematical expression of this model is

$$\ln\left[\frac{Pnijk}{Pnijk-1}\right] = \theta n - \beta i - \alpha j - \tau k$$

where $Pnijk$ = probability of test-taker n being rated k on trait i by rater j, $Pnijk-1$ = probability of test-taker n being rated $k-1$ on trait i by rater j, θn = ability of test-taker n, βi = difficulty of criterion i, αj = severity of rater j, and τk = difficulty of scale category k relative to scale category $k-1$. The same MFRM model was also used to address RQ1 and RQ3.

Second, interactions/bias analysis was run to detect potential interactions between raters and rating scale criteria. The mathematical expression of the model is

$$\ln\left[\frac{Pnijk}{Pnijk-1}\right] = \theta n - \beta i - \alpha j - \tau k - \varphi ji$$

The φji term in the model above denotes each and every combination of an individual rater with each criterion. This model specifies the degree to which rater j's ratings using criterion i deviate from the model expectation. Any statistically significant deviations are indicative of rater bias or differential rater functioning.

Third, to address RQ2, a hybrid model was used by applying the Rating Scale Model (Andrich, 1978) to the scripts and raters facets, and a Partial Credit Model (Wright & Masters, 1982) to the rating scale criteria. The mathematical expression of this hybrid model is

$$\ln\left[\frac{Pnijk}{Pnijk-1}\right] = \theta n - \beta i - \alpha j - \tau ik$$

The τik term in the model above means that FACETS treats individual rating scale criteria as having their own scale structure. Thus, using this hybrid model allowed each criterion in the analytic rating scale to be modelled to have its own scale structure (McNamara et al., 2019).

Further analyses were conducted to provide evidence relevant to RQ3, that is, principal components analysis of residuals using Winsteps v3.81.0 (Linacre, 2015b), and principal axis factoring and correlation analyses using SPSS version 28 (https://www.ibm.com/products/spss-statistics).

Post-rating questionnaire data. To address RQ4, responses to the Likert-type items were analysed using descriptive statistics. Responses to the six open-ended questions—totalling 9209 words (4317 for Rating Scale 1 and 4892 words for Rating Scale 2; contributed by all 20 raters)—were analysed, using Atlas.ti 8.4.5 (https://atlasti.com), for common themes and insights that could elucidate the quantitative results on the ratings data.

Results

Prior to interpreting the MFRM results, data-to-Rasch model fit was examined mainly by inspecting the fit statistics of the three facets involved. Fit statistics, which provide information about the residuals between model expectations and empirical observations, have an expected value of 1.0 and range from 0 to infinity (Linacre, 2020). Table 1 shows that all three facets (examinee, rater, and criterion) for both rating scales had mean infit and outfit statistics of (very close to) 1. Other indicators of data-to-model fit (i.e., unexpected

Table 1. Summary statistics of the three facets.

	Rating Scale 1			Rating Scale 2		
	Examinee	Rater	Criterion	Examinee	Rater	Criterion
Measure						
M	−0.74	.00	.00	.24	.00	.00
SD	1.37	.88	.37	1.24	.79	.42
n	147	20	4	147	20	4
Infit MnSq						
M	0.98	1.00	1.00	0.97	1.00	1.00
SD population	0.46	0.18	0.12	0.42	0.19	0.10
Outfit MnSq						
M	0.97	0.99	0.99	0.97	1.00	1.00
SD population	0.46	0.17	0.10	0.43	0.18	0.10
Separation statistics						
Ratio	3.18	6.32	5.91	2.95	5.76	6.86
Strata	4.57	8.76	8.21	4.26	8.01	9.48
Reliability	.91	.98	.97	.90	.97	.98
Fixed χ^2	1588.5*	796.5*	141.7*	1415.8*	654.4*	191.7*
df	146	19	3	146	19	3

Note: The mean measure of the examinee facet was set to float; that of the rater and rating scale criterion facets was set to 0. *SD*: standard deviation.
*$p < .001$.

responses, fit statistics, estimated discriminations, and inter-rater agreement) also confirmed that the rating data gathered using the two analytic rating scales fit the Rasch measurement model satisfactorily to warrant further interpretations of the results.

RQ1. To what extent do individual raters vary in applying two different analytic rating scales for assessing reading-into-writing?

Rater functioning, rating scale overall functioning, and rater-rating criterion interactions/bias were examined to address RQ1. Table 2 presents abbreviated rater measurement reports for Rating Scales 1 and 2. The severity of the 20 participating raters differed when they were using either Rating Scale 1 or Rating Scale 2. Rater severity variance spread over 2.75 logits from −1.39 to 1.36 logits (*SD* population: 0.88, *SD* sample: 0.90) when using Rating Scale 1, and over 3.38 logits from −1.31 to 2.07 (*SD* population: 0.71, *SD* sample: 0.81) for Rating Scale 2. These results suggest that raters were not homogeneous under both circumstances, yet this lack of consensus is not unexpected (Eckes, 2015; Knoch et al., 2020).

In terms of rating consistency as indicated by the fit statistics from the MFRM results, raters were found to be acceptably consistent when using both Rating Scale 1 and Rating Scale 2. The infit and outfit mean square values of the 20 raters using the two analytic rating scales were well within Linacre's (2002) recommended range of 0.50 to 1.50, with

Table 2. Rater measurement report.

Total count	Measure	Model *SE*	Infit MnSq	ZStd	Exact obs %	Rater
Rating Scale 1						
140	1.36	0.15	1.19	1.5	37.7	16
140	1.31	0.14	1.07	0.6	39.9	14
140	1.30	0.15	0.97	−0.2	36.4	20
140	0.84	0.14	1.12	0.9	41.4	6
140	0.81	0.14	0.67	−3.3	42.9	10
140	0.78	0.14	0.83	−1.4	45.3	5
140	0.54	0.13	1.02	0.2	42.3	13
140	0.31	0.13	0.88	−1.0	41.1	11
140	0.30	0.13	1.44	3.4	33.8	15
140	0.05	0.14	0.82	−1.6	41.1	19
140	−0.06	0.13	0.98	−0.1	46.1	1
140	−0.35	0.13	1.19	1.6	40.2	18
140	−0.40	0.14	0.84	−1.4	47.6	7
140	−0.49	0.14	1.18	1.4	39.7	4
140	−0.52	0.13	0.75	−2.4	45.0	9
140	−0.74	0.13	1.00	0.0	36.8	3
140	−1.17	0.13	1.19	1.6	38.7	8
140	−1.22	0.13	0.93	−0.6	39.1	17
140	−1.25	0.13	0.96	−0.2	37.6	2
140	−1.39	0.14	0.91	−0.7	32.6	12
140.0	0.00	0.14	1.00	−0.1	Mean	(Count: 20)
0.0	0.88	0.00	0.18	1.6	*SD*	(Population)
0.0	0.90	0.01	0.18	1.6	*SD*	(Sample)
Rating Scale 2						
140	2.07	0.14	1.28	2.3	28.8	14
140	1.35	0.14	0.66	−3.4	40.0	20
140	1.03	0.14	0.83	−1.5	38.4	5
140	0.78	0.13	1.04	0.3	38.4	16
140	0.39	0.13	0.99	0.0	38.9	3
140	0.24	0.13	0.86	−1.2	43.8	2
140	−0.01	0.13	0.92	−0.6	39.6	7
140	−0.09	0.14	1.18	1.5	35.6	4
140	−0.10	0.13	0.68	−3.0	43.9	11
140	−0.13	0.13	1.19	1.6	36.9	15
140	−0.15	0.14	1.14	1.1	43.1	18
140	−0.21	0.13	1.08	0.6	39.1	17
140	−0.28	0.14	1.06	0.5	42.1	13
140	−0.30	0.14	1.18	0.5	42.6	19
140	−0.32	0.14	0.92	0.7	44.9	1
140	−0.34	0.13	1.14	0.1	40.3	8
140	−0.72	0.13	0.75	2.3	42.3	9

(Continued)

Table 2. (Continued)

Total count	Measure	Model *SE*	Infit MnSq	ZStd	Exact obs %	Rater
140	−0.77	0.14	0.72	2.6	40.3	12
140	−1.14	0.14	1.04	0.3	41.4	10
140	−1.31	0.14	1.28	0.2	35.8	6
140.0	.00	0.14	1.00	0.1	Mean	(Count: 20)
0.0	0.79	0.00	0.19	0.7	*SD*	(Population)
0.0	0.81	0.00	0.20	0.8	*SD*	(Sample)

Note: SE: standard error; *SD*: standard deviation.

the majority close to 1.0. When using Rating Scale 1, one rater's (Rater 15) infit and outfit statistics (1.44 and 1.42, respectively) were close to the upper limit, suggesting that this rater was rather inconsistent. Another rater (Rater 10) had infit and outfit statistics of 0.67, close to the lower limit, thus leaning towards a slight overfit. When Rating Scale 2 was used, two raters (Raters 20 and 11) also had infit and outfit statistics of .66 and .64, and .68, tending towards a slight overfit. Overfitting raters rate with less variance than expected, possibly indicating a central tendency effect, but this is not detrimental to measurement (McNamara et al., 2019).

Rating scale criterion fit statistics and category statistics showed that both rating scales functioned very well. The infit and outfit mean square values of the criteria in Rating Scale 1 ranged between 0.86 and 1.18 and 0.85 and 1.12, respectively, and for Rating Scale 2 between 0.82 and 1.09 and 0.84 and 1.08, respectively (see Table 3). Examination of rating scale category statistics showed that both Rating Scale 1 and Rating Scale 2 met all the seven requirements for optimum functioning outlined by McNamara et al. (2019) (see Table 4): (1) there were at least 10 observations per score category to ensure the stability of the measurement results; (2) average measures increased by approximately the same value at each score category; (3) average measures increased monotonically; (4) the frequency of data points in each score category formed a reasonably smooth distribution with only one peak; (5) average measures were close to expected measures; (6) mean square values were all below 2.0; and (7) Rasch–Andrich thresholds increased monotonically at each score point by between 1.4 and 5 logits.

Analyses of rater and rating scale criterion interactions/bias provide detailed information about individual raters' severity/leniency measures against every criterion in the rating scale. Statistically significant interactions suggest that raters interpret the difficulty of certain criteria either significantly lower or higher than the group (Wind & Engelhard, 2013). Table 5 shows eight statistically significant interactions between raters and Rating Scale 1 criteria. Interactions with t values above 2.0 suggest that raters were rating more leniently than the model expectation; conversely, t values below −2.0 indicate that raters were rating more severely than expected by the model. There is no single criterion that caused persistent bias among raters, but the *Language Control* criterion appeared to have extreme interactions with Rater 15 (extremely severe) and Rater 18 (extremely lenient). In an operational rating condition, this information could be used to provide individualized feedback to the raters concerned (see Knoch, 2011).

Table 3. Rating scale criterion measurement report.

Total score	Total count	Observed average	Fair average	Measure	Model *SE*	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. discrim.	Corr. PtMea	Corr. PtExp	*n*	Criterion
Rating Scale 1														
1407	700	2.01	1.95	.55	.06	1.18	3.3	1.12	2.0	0.84	.71	.69	1	RW
1530	700	2.19	2.15	.08	.06	0.86	−2.9	0.85	−2.9	1.18	.74	.70	2	TF
1604	700	2.29	2.27	−.19	.06	0.99	−0.1	1.01	0.2	0.97	.66	.70	4	LC
1671	700	2.39	2.38	−.44	.06	0.96	−0.8	0.97	−0.5	1.03	.68	.70	3	OS
1553.0	700.0	2.22	2.19	.00	.06	1.00	−0.1	0.99	−0.3		.70		Mean	(Count: 4)
97.9	0.0	.14	.16	.37	.00	0.12	2.3	0.10	1.8		.03		*SD*	(Population)
113.1	0.0	.16	.18	.42	.00	0.14	2.6	0.11	2.1		.04		*SD*	(Sample)
Rating Scale 2														
1666	700	2.38	2.38	.52	.06	1.09	1.7	1.08	1.5	0.92	.70	.67	1	SU
1733	700	2.48	2.48	.28	.06	0.82	−3.7	0.84	−3.4	1.21	.72	.67	3	DI
1885	700	2.69	2.72	−.28	.06	1.05	0.9	1.06	1.1	0.94	.64	.67	2	OG
1953	700	2.79	2.83	−.53	.06	1.02	0.4	1.03	0.5	0.95	.61	.67	4	LU
1809.3	700.0	2.58	2.60	.00	.06	1.00	−0.1	1.00	0.0		.67		Mean	(Count: 4)
114.8	.0	.16	.18	.42	.00	0.10	2.1	0.10	2.0		.05		*SD*	(Population)
132.6	.0	.19	.21	.49	.00	0.11	2.5	0.11	2.3		.05		*SD*	(Sample)

Note: SE: standard error; RW: Reading for Writing; TF: Task Fulfilment; OS: Organization and Structure; LC: Language Control; *SD*: standard deviation; SU: Source Use; OG: Organization; DI: Development of Ideas.

Table 4. Rating scale category statistics.

	Data			Quality control			Rasch–Andrich thresholds		Expectation measure at		Most probable from	Rasch–Thurstone thresholds	Cat peak prob
	Score	Category counts	%	Ave. mea.	Exp. mea.	Outfit MnSq	Mea.	*SE*	Cat.	−0.5			
Rating Scale 1	1	625	22%	−2.38	−2.30	0.9			(−3.40)		Low	Low	100%
	2	1150	41%	−0.99	−1.03	0.9	−2.26	.06	−1.17	−2.48	−2.26	−2.35	60%
	3	813	29%	0.35	0.27	1.0	−0.04	.05	1.16	−0.01	−0.04	−0.03	61%
	4	212	8%	1.39	1.62	1.1	2.30	.08	(3.44)	2.50	2.30	2.37	100%
Rating Scale 2	1	284	10%	−1.84	−1.71	0.9			(−3.47)		Low	Low	100%
	2	994	36%	−0.43	−0.47	1.0	−2.34	.07	−1.19	−2.53	−2.34	−2.42	61%
	3	1123	40%	0.72	0.70	1.0	−0.01	.05	1.19	−0.00	−0.01	−0.00	61%
	4	399	14%	1.88	1.95	1.0	2.34	.06	(3.49)	2.55	2.34	2.42	100%

Note: SE: standard error.

Table 5. Rating Scale 1 rater-rating criterion interaction/bias.

Obs. score	Exp. score	Obs. count	Obs-exp ave.	Bias size	Model *SE*	*t*	*f*	Prob.	Infit MnSq	Outfit MnSq	Rater		Rating Scale 1 criterion	
											Nu	Mea.	Crit.	Mea.
100	81.92	35	.52	1.29	.27	4.75	34	.0000	0.8	0.8	18	−0.35	LC	−.19
89	76.74	35	.35	0.90	.27	3.33	34	.0021	1.1	1.0	4	−0.49	RW	.55
87	74.88	35	.35	0.85	.26	3.22	34	.0028	0.8	0.7	15	0.30	TF	.08
75	65.44	35	.27	0.72	.27	2.68	34	.0013	0.9	1.0	16	1.36	OS	−.44
82	90.15	35	−.23	−0.60	.27	−2.21	34	.0340	0.7	0.7	4	−0.49	OS	−.44
62	71.65	35	−.28	−0.76	.29	−2.62	34	.0130	1.4	1.2	18	−0.35	RW	.55
56	67.38	35	−.33	−0.94	.30	−3.10	34	.0039	1.0	0.9	1	−0.06	RW	.55
62	78.76	35	−.48	−1.25	.29	−4.36	34	.0001	1.3	1.2	15	0.30	LC	−.19

Note: SE: standard error; RW: Reading for Writing; TF: Task Fulfilment; OS: Organization and Structure; LC: Language Control.

As can be seen from Table 6, there were quite a few statistically significant interactions between raters and Rating Scale 2 criteria. One particularly important observation concerns the *Language Use* criterion, which appeared to cause the highest number of interactions (7) of all the criteria. Furthermore, some raters displayed differential severity and/or leniency against certain criteria. Myford and Wolfe (2004) noted that differential severity against an easy criterion and/or differential leniency against a difficult criterion might indicate an individual halo effect. For example, Rater 4 exhibited a statistically significant differential severity against the easiest criterion (*Language Use*) and a statistically significant differential leniency against the most difficult criterion (*Source Use*). Rater 4's overall severity measure (−.09) was very close to the group average (.00), but this masks this rater's differential leniency against the *Source Use* criterion and differential severity against the *Language Use* criterion in Rating Scale 2, as these two differential severity and leniency measures cancel each other out.

RQ2. How do raters apply the two analytic rating scales across individual rating criteria?

To address RQ2, results from three different analyses are presented: (1) hybrid MFRM analysis allowing the inspection of each criterion within each rating scale to be modelled separately, (2) category statistics of individual criteria, and (3) individual criterion discrimination power.

The results of the hybrid MFRM analysis are illustrated in the Wright maps in Figures 6 and 7, which display the relationships between test-taker ability, rater severity/leniency, and rating scale criterion difficulty with the individual criteria modelled to have their own scale structure. It is evident that raters used the scale structures (Scores 1–4) for the individual rating criteria in each scale differently. In Rating Scale 1, the *Reading for Writing* criterion appeared to have a distinct scale structure from the other three criteria, indicated by the locations of the category thresholds marked by the horizontal dashes in Figure 6. Likewise, the *Source Use* criterion in Rating Scale 2 displayed a similar observation (see Figure 7); these two criteria exclusively representing the reading aspect of reading-into-writing seemed to have a narrower scale structure. Therefore, it appeared less likely for examinees at the lower end of the logit scale to achieve a score of 2 in *Reading for Writing* in Rating Scale 1 or *Source Use* in Rating Scale 2 than to achieve the same score in the other three criteria in the respective rating scales. Conversely, it was more likely for examinees at the higher end of the logit scale to achieve a score of 4 in *Reading for Writing* in Rating Scale 1 or *Source Use* in Rating Scale 2 than to achieve the same score in the other three criteria in the respective rating scales.

In terms of category statistics, the individual rating criteria in each rating scale satisfied almost all requirements for optimum functioning of rating scales (see Table 7). This provides more evidence that the criteria in both rating scales were well functioning.

Individual criterion discrimination, or the ability of each individual criterion within an analytic rating scale to discriminate examinees, was also examined. Less discriminating criteria may indicate raters' difficulty in using the criteria (Knoch et al., 2020). As shown in Table 8, *Reading for Writing* was the least discriminating criterion in

Table 6. Rating Scale 2 rater-rating criterion interactions/bias.

Obs. score	Exp. score	Obs. count	Obs-exp ave.	Bias size	Model *SE*	*t*	*f*	Prob.	Infit MnSq	Outfit MnSq	Rater		Rating Scale 2 criterion	
											Nu	Mea.	Crit.	Mea.
103	84.84	35	.52	1.34	.28	4.84	34	.0000	0.8	0.8	4	−.09	SU	.52
114	98.02	35	.46	1.28	.30	4.30	34	.0001	1.6	1.6	18	−.15	LU	−.53
116	104.92	35	.32	.90	.29	3.04	34	.0045	1.0	1.1	13	−.28	LU	−.53
114	104.74	35	.26	.72	.29	2.51	34	.0170	0.5	0.5	9	−.72	LU	−.53
99	89.48	35	.27	.70	.27	2.55	34	.0154	0.8	0.8	19	−.30	DI	.28
97	88.02	35	.26	.64	.27	2.38	34	.0230	0.6	0.5	11	−.10	DI	.28
85	76.76	35	.24	.60	.27	2.23	34	.0326	0.9	0.9	3	.39	SU	.52
94	101.58	35	−.22	−.55	.27	−2.07	34	.0465	1.2	1.2	13	−.28	OG	−.28
88	95.84	35	−.22	−.56	.27	−2.10	34	.0435	0.5	0.5	11	−.10	OG	−.28
102	109.71	35	−.22	−.59	.27	−2.16	34	.0378	0.5	0.7	12	−.77	LU	−.53
75	83.58	35	−.25	−.63	.27	−2.30	34	.0276	0.7	0.7	18	−.15	SU	.52
91	100.49	35	−.27	−.70	.27	−2.59	34	.0139	0.9	0.9	19	−.30	LU	−.53
92	101.87	35	−.28	−.72	.27	−2.68	34	.0112	1.2	1.2	15	−.13	LU	−.53
91	101.00	35	−.29	−.73	.27	−2.72	34	.0101	1.4	1.4	6	−1.31	SU	.52
88	99.17	35	−.32	−.82	.27	−3.05	34	.0045	0.8	0.8	4	−.09	LU	−.53

Note: SE: standard error; SU: Source Use; OG: Organization; DI: Development of Ideas; LU: Language Use.

```
+------+------------+------------+-------------+-----+-----+-----+-----+
|Logit |+Script     |-Rater      |-Criterion   | RW  | TF  | OS  | LC  |
+------+------------+------------+-------------+-----+-----+-----+-----+
+   3  + High       + Severe     + Difficult   + (4) + (4) + (4) + (4) +
|      | .          |            |             |     |     | --- | --- |
|      |            |            |             |     |     |     |     |
|      | *          |            |             |     | --- |     |     |
|      |            |            |             |     |     |     |     |
+   2  + .          +            +             + --- +     +     +     +
|      | *.         |            |             |     |     |     |     |
|      | *          |            |             |     |     |     |     |
|      | *.         | 14  16  20 |             |     |     |  3  |     |
|      | *.         |            |             |     |  3  |     |  3  |
+   1  + **.        +            +             +     +     +     +     +
|      | *.         | 10  5   6  |             |  3  |     |     |     |
|      | *.         | 13         |             |     |     |     |     |
|      | ***.       | 11         | RW          |     |     |     |     |
|      | *.         | 15         |             |     |     |     |     |
*   0  * ***.       * 1   19     * TF  LC      * --- * --- * --- * --- *
|      | ****       |            |             |     |     |     |     |
|      | **         | 18  4   7  | OS          |     |     |     |     |
|      | *********. | 9          |             |     |     |     |     |
|      | ****.      | 3          |             |     |     |     |     |
+  -1  + ****       +            +             +  2  +     +     +     +
|      | ****.      | 17  2   8  |             |     |  2  |  2  |     |
|      | ****.      | 12         |             |     |     |     |  2  |
|      | ****.      |            |             |     |     |     |     |
|      | **.        |            |             |     |     |     |     |
+  -2  + **         +            +             + --- +     +     +     +
|      | **         |            |             |     |     |     |     |
|      | **         |            |             |     |     |     |     |
|      | **         |            |             |     | --- |     |     |
|      | .          |            |             |     |     | --- | --- |
+  -3  + .          +            +             +     +     +     +     +
|      | .          |            |             |     |     |     |     |
|      | .          |            |             |     |     |     |     |
|      | .          |            |             |     |     |     |     |
|      | .          |            |             |     |     |     |     |
+  -4  +            +            +             +     +     +     +     +
|      |            |            |             |     |     |     |     |
|      | .          |            |             |     |     |     |     |
|      |            |            |             |     |     |     |     |
|      |            |            |             |     |     |     |     |
+  -5  +            +            +             +     +     +     +     +
|      |            |            |             |     |     |     |     |
|      |            |            |             |     |     |     |     |
|      |            |            |             |     |     |     |     |
|      | .          |            |             |     |     |     |     |
+  -6  + Low        + Lenient    + Easy        + (1) + (1) + (1) + (1) +
+------+------------+------------+-------------+-----+-----+-----+-----+
|Logit |+Script     |-Rater      |-Criterion   | RW  | TF  | OS  | LC  |
+------+------------+------------+-------------+-----+-----+-----+-----+
```

Figure 6. Rating Scale 1 Wright map from the many-facet partial credit analysis.
Note: Each star in the second column represents two examinees, and a dot represents one examinee. The horizontal dashes in the last four columns indicate the category thresholds. RW = Reading for Writing; TF = Task Fulfilment; OS = Organization and Structure; LC = Language Control.

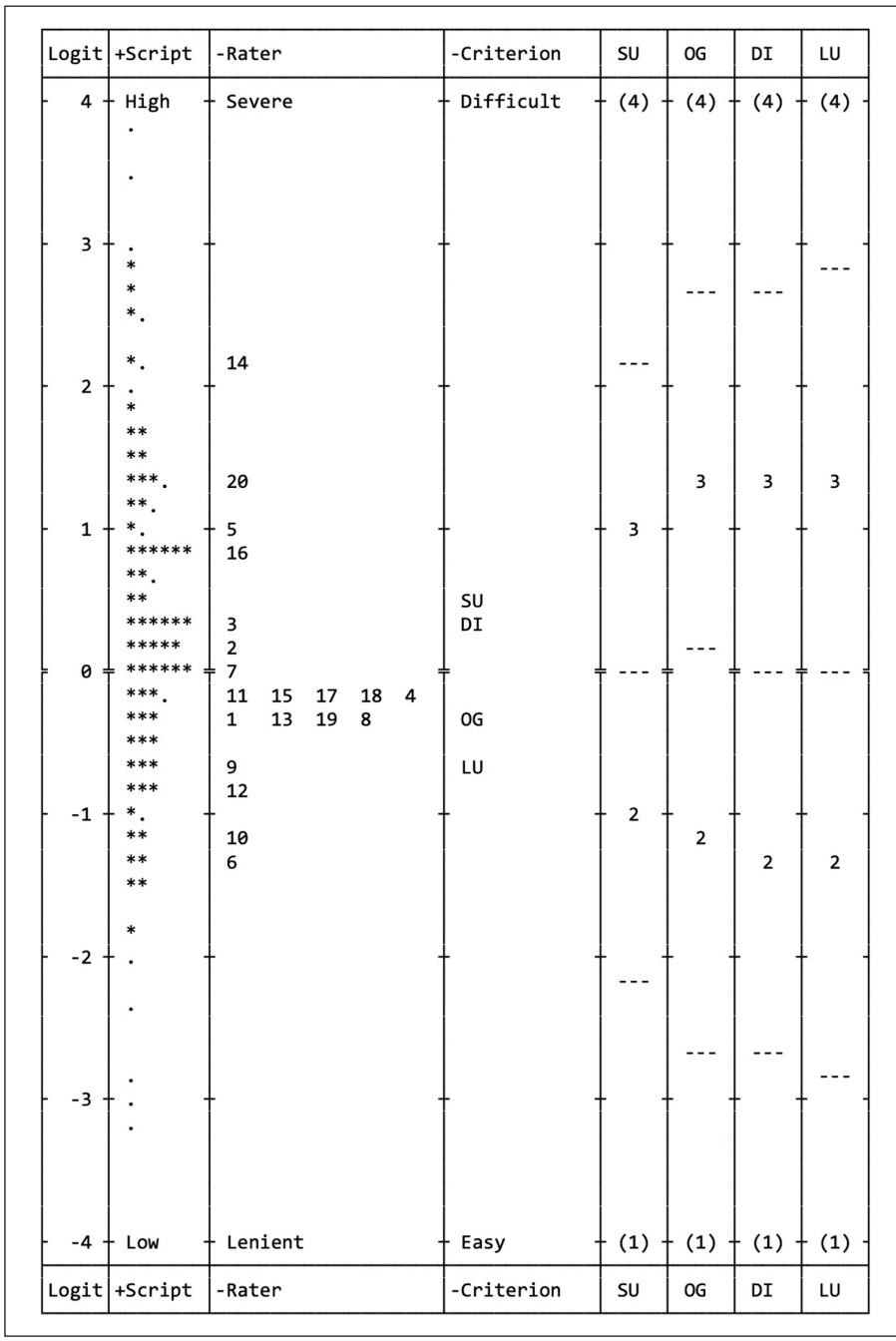


Figure 7. Rating Scale 2 Wright map from the many-facet partial credit analysis.
Note: Each star in the second column represents two examinees, and a dot represents one examinee. The horizontal dashes in the last four columns indicate the category thresholds. SU: Source Use; OG: Organization; DI: Development of Ideas; LU: Language Use.

Table 7. Rating Scales 1 and 2 summary of category statistics.

Requirement	Rating Scale 1				Rating Scale 2			
	Reading for Writing	Task Fulfilment	Organization and Structure	Language Control	Source Use	Organization	Development of Ideas	Language Use
>10 observations per category	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Average measures advance by approximately the same value	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Average measures increase monotonically and are not disordered	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Smooth distribution of frequency	No[a]	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Average measures ≅ average expected measures	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean squares > 2	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Rasch–Andrich thresholds increase by between 1.4 and 5 logits	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

[a]Score 1 was the most frequently used (36%).

Table 8. Sub-scale discrimination.

	Criterion	Range of examinee ability (logit)[a]
Rating Scale 1	Reading for Writing	9.24
	Language Control	10.06
	Organization and Structure	10.75
	Task Fulfilment	11.17
Rating Scale 2	Language Use	9.17
	Development of Ideas	9.71
	Organization	9.79
	Source Use	11.48

[a]This was calculated by running separate many-facet Rasch measurement analyses for each criterion.

Rating Scale 1, while its counterpart, *Source Use*, was the most discriminating criterion in Rating Scale 2.

RQ3. To what extent do each analytic scale's ratings provide unique but complementary information?

This question essentially investigates the extent to which the four criteria within each rating scale measure distinct aspects (*multidimensionality*) while working together to measure one underlying construct (*unidimensionality*). These features were examined using five indicators: (1) MFRM analysis of criterion fit statistics, (2) MFRM analysis of group-level halo effects (lack thereof), (3) principal components analysis of residuals, (4) correlation matrix, and (5) principal axis factoring.

The rating scale criterion fit statistics shown in Table 3 indicated that the criteria within each rating scale meet the expectations for unidimensional measurement required for MFRM, meaning that together they can be considered to form a single cohesive pattern of observations. This in turn can indicate that the four criteria measure one underlying construct of reading-into-writing.

Investigation of group-level halo effect, or more precisely the lack thereof, for the purpose of this research, could provide evidence that raters are able to distinguish the criteria within an analytic rating scale, and this can be done by examining the criterion difficulty measures and several indicators in the rating scale criterion summary statistics (Myford & Wolfe, 2004). Halo effect refers to "a rater's tendency to assign ratees similar ratings in conceptually distinct traits" (Myford & Wolfe, 2004, p. 209). The criterion difficulty measures of Rating Scale 1 and Rating Scale 2 in Table 3 show that the four criteria in each rating scale had different difficulty levels. Furthermore, as shown in Table 1, the significant homogeneity index (fixed χ^2) and its degrees of freedom reject the null hypothesis that the different criteria in both Rating Scale 1 and Rating Scale 2 are of the same difficulty levels. The high reliability coefficients (.97 for Rating Scale 1 and .98 for Rating Scale 2) provide further evidence that the four criteria were reliably separate.

Following Knoch et al. (2020), this study also used principal components analysis of residuals to examine the unidimensionality of the two analytic rating scales. The

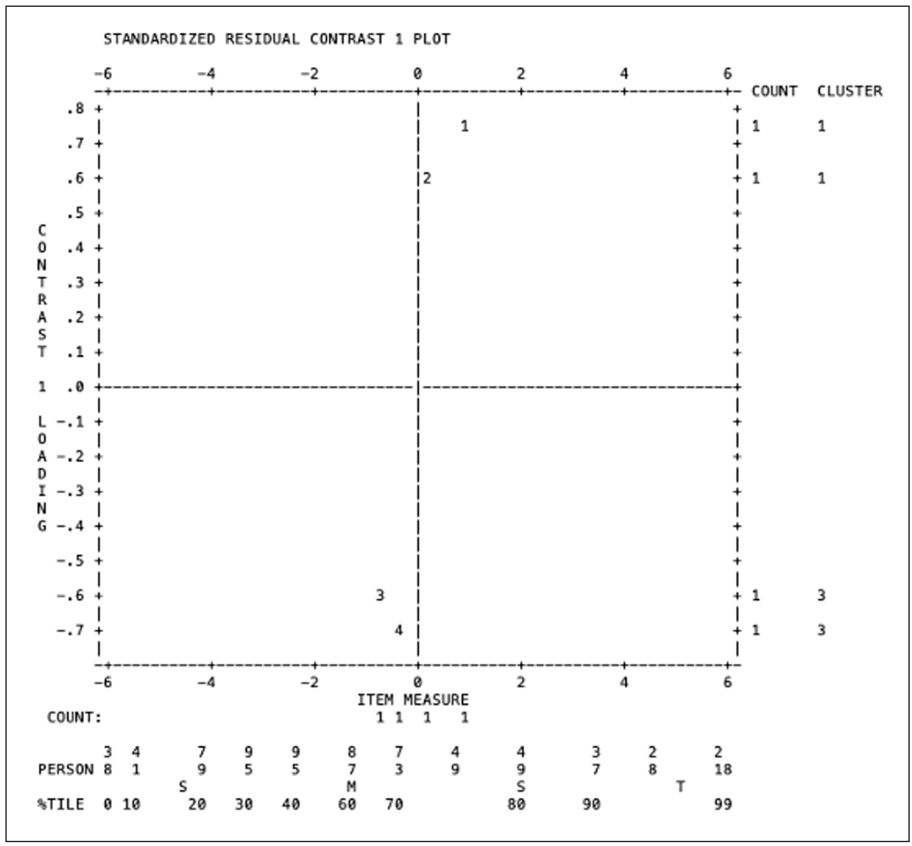


Figure 8. PCAR for Rating Scale 1.
Note: PCAR = principal components analysis of residuals. 1 = Reading for Writing; 2 = Task Fulfilment; 3 = Organization and Structure; 4 = Language Control.

unexplained variances in the first contrast (i.e., the eigenvalue of the first principal component residuals) in both Rating Scale 1 and Rating Scale 2 were 1.8 and 1.9, respectively. A value below 2.0 indicates the absence of a secondary dimension (Linacre, 2020), which means that the four criteria within each rating scale worked in harmony to measure one unified construct. However, it is important to note that two "strands" were identified in both rating scales (Figures 8 and 9), and it can be inferred that Strand 1 is strongly related to content, while Strand 2 is more oriented towards organization and language aspects.

Correlation matrices are often used to examine the interrelationships among criteria within an analytic rating scale (e.g., Knoch et al., 2020; Xi & Mollaun, 2006). As shown in Tables 9 and 10, the four criteria within each rating scale are strongly intercorrelated, suggesting that they are related, but not overly strong to make any of the criteria redundant.

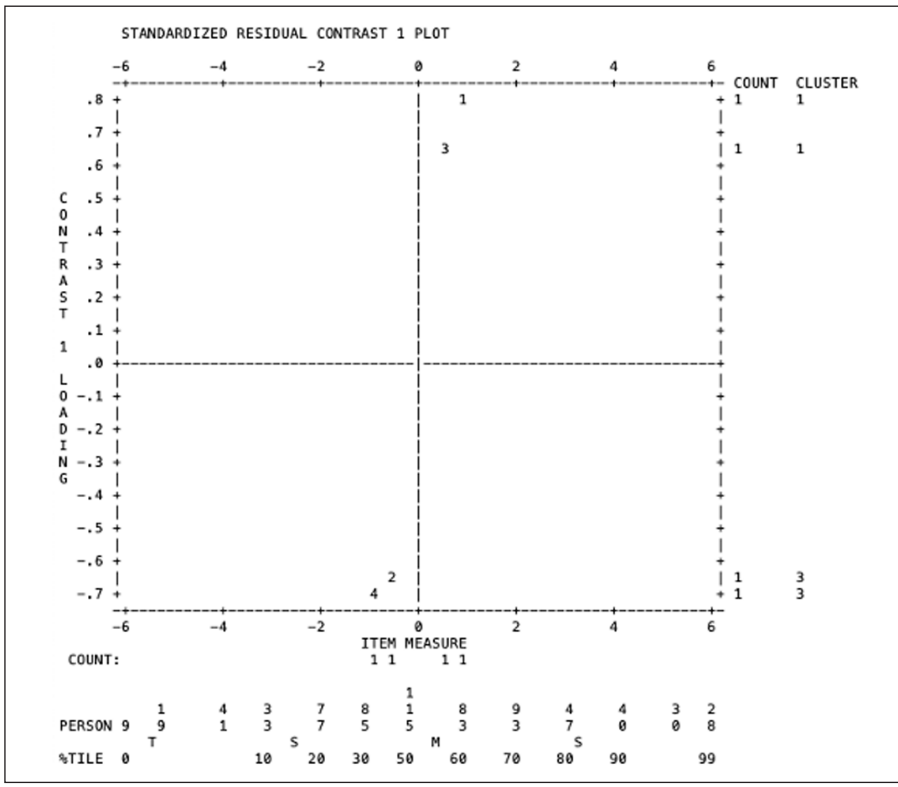


Figure 9. PCAR for Rating Scale 2.
Note: PCAR = principal components analysis of residuals. 1 = Source Use; 2 = Organization; 3 = Development of Ideas; 4 = Language Use.

Table 9. Correlations among Rating Scale 1 criteria.

Criterion	RW	TF	OS	LC
Reading for Writing (RW)	–			
Task Fulfilment (TF)	.747	–		
Organization and Structure OS)	.563	.652	–	
Language Control (LC)	.516	.563	.625	–

Table 10. Correlations among Rating Scale 2 criteria.

Criterion	SU	OG	DI	LU
Source Use (SU)	–			
Organization (OG)	.535	–		
Development of Ideas (DI)	.737	.591	–	
Language Use (LU)	.437	.587	.542	–

Table 11. Summary of analysis for RQ3.

Analysis	Unidimensionality	Multidimensionality
MFRM—criterion fit statistics	✓	
MFRM—absence of group-level halo effect		✓
Principal components analysis of residuals	✓	
Correlations	✓	✓
Principal axis factoring	✓	

Note: MFRM: many-facet Rasch measurement.

Principal axis factoring, a type of exploratory factor analysis, is also useful to examine the unidimensionality of analytic rating scales (Knoch et al., 2020). The correlation matrix determinant, Kaiser–Meyer–Olkin (KMO) measure of sample adequacy, and Bartlett's test of sphericity result (.139, .780, and sig. at .000, respectively, for Rating Scale 1; .172, .757, and sig. at .000, respectively, for Rating Scale 2) indicate that the data were suitable for factor analysis. For both rating scales, only one factor was detected. For Rating Scale 1, the factor had an eigenvalue of 2.836, accounting for 70.991% of the total variance, and for Rating Scale 2, the factor had an eigen value of 2.720, accounting for 67.990% of the total variance. This provides further evidence that the four criteria within each rating scale investigated work harmoniously to measure one underlying construct, reading-into-writing.

The five different indicators addressing RQ3 are summarized in Table 11.

RQ4. What are raters' perceptions of the two analytic rating scales?

Descriptive statistics of raters' responses on the 15 Likert-type post-rating questionnaire items are presented in Table 12. It is evident that both rating scales were generally well received by the raters, indicated by a grand mean of above 3.00 for both rating scales (M=3.77 for Rating Scale 1; M=3.59 for Rating Scale 2). For Rating Scale 1, the *Language Control* (M=4.07) and *Organization and Structure* (M=3.87) criteria received very positive responses, indicating that raters found these criteria to have clear descriptors, help them discriminate among examinee performance levels, and help them rate reliably and consistently. Although the *Reading for Writing* (M=3.72) and *Task Fulfilment* (M=3.52) criteria were deemed to have clear descriptors, they were perceived comparatively less helpful than the other two criteria in terms of facilitating raters to mark reliably and consistently. For Rating Scale 2, the *Source Use* (M=3.82), *Organization* (M=3.72), and *Language Use* (M=3.85) criteria were perceived favourably by the raters. *Development of Ideas*, however, received a comparatively less favourable response (M=3.18).

In raters' responses to the open-ended questions, two general themes were identified: (1) rating scale dimensionality and (2) facilitating and/or hindering features of the rating scale. Raters held mixed views on the dimensionality of both rating scales. While in general, raters reported that they were able to distinguish the individual criteria within

Table 12. Raters' responses to the perception Likert-style statements.

Statement	%					*M* (*SD*)
	1	2	3	4	5	
Rating Scale 1						
The rating scale is practical to use.	5%	5%	15%	60%	15%	3.75 (0.97)
The descriptors under *Reading for Writing*						
• are clear.	0%	5%	10%	70%	15%	3.95 (0.69)
• helped me distinguish different performance levels.	0%	15%	20%	50%	15%	3.65 (0.93)
• helped me mark the performances reliably and consistently.	0%	10%	40%	35%	15%	3.55 (0.89)
The descriptors under *Task Fulfilment*						
• are clear.	5%	5%	10%	70%	10%	3.75 (0.91)
• helped me distinguish different performance levels.	0%	20%	20%	60%	0%	3.40 (0.82)
• helped me mark the performances reliably and consistently.	0%	15%	35%	45%	5%	3.40 (0.82)
The descriptors under *Organization and Structure*						
• are clear.	0%	0%	5%	75%	20%	4.15 (0.49)
• helped me distinguish different performance levels.	0%	5%	20%	65%	10%	3.80 (0.70)
• helped me mark the performances reliably and consistently.	0%	5%	35%	50%	10%	3.65 (0.75)
The descriptors under *Language Control*						
• are clear.	0%	0%	0%	75%	25%	4.25 (0.44)
• helped me distinguish different performance levels.	0%	0%	20%	60%	20%	4 .00 (0.65)
• helped me mark the performances reliably and consistently.	0%	0%	25%	55%	20%	3.95 (0.69)
With regard to test-takers' use of information from the source texts, the rating scale provides adequate guidance for me to make scoring decisions reliably and consistently.	5%	10%	30%	45%	10%	3.45 (1.00)
In general, the rating scale provides useful guidance for me to make scoring decisions.	5%	0%	10%	70%	15%	3.90 (0.85)
Rating Scale 2						
The rating scale is practical to use.	5%	10%	25%	45%	15%	3.55 (1.05)
The descriptors under *Source Use*						
• are clear.	0%	5%	15%	65%	15%	3.90 (0.72)
• helped me distinguish different performance levels.	0%	15%	5%	65%	15%	3.80 (0.89)
• helped me mark the performances reliably and consistently.	0%	10%	20%	55%	15%	3.75 (0.85)

(Continued)

Table 12. (Continued)

Statement	%					*M* (*SD*)
	1	2	3	4	5	
The descriptors under *Organization*						
• are clear.	0%	15%	10%	55%	20%	3.80 (0.95)
• helped me distinguish different performance levels.	0%	10%	15%	65%	10%	3.75 (0.79)
• helped me mark the performances reliably and consistently.	0%	10%	30%	50%	10%	3.60 (0.82)
The descriptors under *Development of Ideas*						
• are clear.	5%	20%	25%	40%	10%	3.30 (1.08)
• helped me distinguish different performance levels.	0%	25%	30%	45%	0%	3.20 (0.83)
• helped me mark the performances reliably and consistently.	5%	20%	40%	35%	0%	3.05 (0.89)
The descriptors under *Language Use*						
• are clear.	0%	5%	25%	40%	30%	3.95 (0.89)
• helped me distinguish different performance levels.	0%	5%	25%	50%	20%	3.85 (0.81)
• helped me mark the performances reliably and consistently.	0%	5%	30%	50%	15%	3.75 (0.79)
With regard to test-takers' use of information from the source texts, the rating scale provides adequate guidance for me to make scoring decisions reliably and consistently.	10%	15%	30%	35%	10%	3.20 (1.15)
In general, the rating scale provides useful guidance for me to make scoring decisions.	5%	20%	10%	55%	10%	3.45 (1.10)

Note: 1 = *strongly disagree*, 2 = *disagree*, 3 = *neither agree nor disagree*, 4 = *agree*, 5 = *strongly disagree*.

each rating scale, they also reported perceiving overlaps between two or more criteria. On the distinguishability among the criteria in Rating Scale 1, one rater said,

> Yes, the criteria were adequately distinguishable. This was mainly due to the detailed descriptions/features provided under each criterion, which made the scale rather clear.

Similarly, commenting on the distinguishability of the criteria in Rating Scale 2, a rater remarked "Yes the criteria were clearly defined and each category was clear as to what area I was looking at when marking."

However, for Rating Scale 1, 12 raters perceived overlaps between *Reading for Writing* and *Task Fulfilment*, and between *Task Fulfilment* and *Organization and Structure.* The former is illustrated in this comment:

> Mostly, though Reading for Writing and Task Fulfilment were a little more difficult to distinguish i.e. "adaptation of content to suit the purpose for writing" may also be related to achieving the overall communicative aim. I think separating these two is generally very challenging.

The impact of perceived overlaps on rating decisions appeared to vary. For instance, one rater felt that the perceived overlap between *Reading for Writing* and *Task Fulfilment* could result in score interdependence between the two:

> I thought that Reading for Writing and Task achievement were very closely linked, and a low score on one meant that the other must also have a low score.

However, that was not always the case, as illustrated by this quote:

> At times, "Task Fulfilment" and "Organisation and Structure" seemed to have a little overlap. I struggled to distinguish between them on a few occasions, but this didn't cause me much hindrance.

Similarly, for Rating Scale 2, 14 raters reported perceiving overlaps mainly between *Source Use* and *Development of Ideas*, and sometimes also with *Language Use*, as these three criteria contain descriptors relating to use of source information. The quote below illustrates this perceived overlap and how the rater tried to define the distinction between the two criteria, while also acknowledging the challenge of this distinction.

> There appears to be excessive overlap between the descriptors for source use and development of ideas. Perhaps the source use criteria could be interpreted as what points/how many points were mentioned from the source text and the development of ideas as how well these ideas supported the points made in the essay. However, this distinction would only work with high level candidates as candidates who hardly refer to the sources can not use ideas from the text which they have not mentioned to support points, making the criteria irrelevant.

Furthermore, two raters also stated that lack of source use could impact the evaluation of test-takers' *Language Use*, as one rater commented:

> The language use was particularly difficult to mark if they had not really referred to the sources.

In terms of facilitating and/or hindering features, both rating scales received positive remarks on focusing raters to assess the reading and source use aspects, as illustrated in these quotes:

> As I said the reference to having points that show the reading for understanding has been completed [. . .] is "right on the money" as far as assessment is concerned. It serves as a tick list for the assessor. (Rating Scale 1)

> I found the constructs under "Source use" useful, particularly the requirement that source content should be integrated, accurately presented and effective (not just included). (Rating Scale 2)

The use of adjectives such as excellent, good, appropriate, and adequate in both rating scales' descriptors drew mixed reviews. While these adjectives are open to interpretations and therefore make rating more subjective and less agreeable, as reported by the raters, they actually helped them remain consistent within themselves.

> [The] grading within categories (excellent; good; acceptable; poor) helped me being consistent about rating. (Rating Scale 1)

> I think grading such as "adequate/ good/ some / few" in the criteria section help me to draw a clear line between performances. It was very helpful [. . .] (Rating Scale 2)

The length and density of descriptors were also valued differently by the raters. The rather lengthy descriptors in Rating Scale 1 were appreciated by 12 raters for providing clear guidance, yet viewed by six raters as at times *overanalytical* and impractical as increasing rating time. For example, one rater stated,

> It takes longer to familiarise [. . .] and internalize [. . .]. I repeatedly had to double-check descriptors during the rating process for accuracy and consistency [. . .]

Rating Scale 2 was praised by five raters for being *brief, simple*, and *straightforward* and therefore *user-friendly* and *time-saving*, while criticized by eight raters for lacking detail, which according to one rater, could potentially lead to "impression [marking]."

Discussion

The purpose of this study was to investigate the potential effects of two different ways of reading-into-writing construct operationalization in analytic rating scales on raters' rating reliability and consistency. The results of both the statistical and qualitative analyses show that both ways of construct operationalization, as reflected in the two analytic rating scales used in this study, result in reliable and consistent ratings, including of reading-related and source use aspects. Participating raters also reported that both rating scales prompted them to mark reading and source use aspects. These findings echo earlier findings of studies on individual analytic rating scales (Chan et al., 2015; Ohta et al., 2018) and provide further support for the use of analytic rating scales to score reading-into-writing task performances.

The quantitative rating data analyses showed that both rating scales functioned optimally and that raters were generally consistent with themselves despite their varying severity levels. These findings were further supported by the questionnaire findings which indicated that raters were generally very positive towards either scale. The intra-rater consistency and variable severity/leniency levels could be partially explained by raters' mixed views on the quantifiers and adjectives in each rating scale's descriptors; while these features were useful to maintain intra-rater consistency, the raters felt that they could be interpreted differently, thereby potentially contributing to severity/leniency variability.

Closer examination of raters and rating criteria interactions/bias revealed that Rating Scale 2, particularly the *Language Use* criterion, prompted more interactions, with some raters exhibiting differential severity and leniency against this criterion. Compared to the *Language Control* criterion in Rating Scale 1, which only interacted statistically significantly with two raters, the *Language Use* criterion in Rating Scale 2 did so with eight raters. As noted by Wind and Engelhard (2013), statistically significant interactions

between raters and rating scale criteria are indicative of raters having different interpretations of the difficulty level of certain criteria. A plausible explanation for this is the inclusion of source use aspects in the descriptors under the *Language Use* criterion in Rating Scale 2 (e.g., *language errors may misrepresent the source ideas*). Indeed, some raters indicated in the questionnaire that failure to integrate adequate source use information in test-takers' writing could influence the evaluation of *Language Use*. While raters are clearly familiar with evaluating language aspects in written performances, they may not be accustomed to evaluating language aspects in relation to how they are used to represent ideas from source texts. Consequently, more raters were likely to have resorted to their own interpretation of part of the *Language Use* descriptors in Rating Scale 2, hence their differential leniency/severity against this criterion. Reference to source use might therefore best be avoided in descriptors of a criterion assessing language aspects in analytic rating scales for reading-into-writing.

Further examination of the individual criteria in each rating scale found that those criteria exclusively representing the reading-related and source use aspects (i.e., *Reading for Writing* in Rating Scale 1 and *Source Use* in Rating Scale 2) appeared to be the most difficult and to have a different scale structure compared to the other criteria in their respective scale. Possible reasons for the difficulty of reading-related aspects might be due to test-takers' lack of or lower familiarity with the relatively new task type of integrated tasks (Chan et al., 2015) or their proficiency level being below the *threshold* for this task type (Cumming, 2014). However, these are less plausible in the case of ISE II test-takers, who have typically followed the associated curriculum and been entered for this specific test level. The different scale structure which the two reading-related criteria were found to have in this study suggest that raters interpreted the difficulty of these criteria differently from others in the rating scale. Potentially, raters lacked familiarity with rating scales containing a "novel" criterion on reading-related and source use aspects; however, no raters raised this issue in the questionnaire. While difficulty differences between analytic criteria are actually not uncommon (Wind, 2020), this warrants further investigation.

The *Reading for Writing* criterion was also found to be the least discriminating of all the criteria in Rating Scale 1, and the lowest score of 1 was the most frequently awarded score of all the four score categories for this criterion. One obvious reason seems test-takers' low performance in this criterion. However, a closer look at the descriptors in score 2 might offer an alternative explanation. To score a 2 in *Reading for Writing*, a test-taker needs to demonstrate "full and accurate understanding of more than half of the source materials." This requirement might have prompted raters to count the number of ideas that test-takers have fully and accurately included in their response, and failure to do so automatically resulted in a score of 1. Raters' questionnaire responses also suggested that the descriptors under *Reading for Writing*, along with those under *Task Fulfilment*, were not as helpful as those under *Organization and Structure* and *Language Control* in terms of facilitating raters to distinguish between performance levels and to mark reliably and consistently. In Rating Scale 2, on the contrary, the *Source Use* criterion was found to be the most discriminating of all criteria, and this quantitative result was also in line with raters' questionnaire responses that the *Source Use* descriptors were the clearest and most helpful.

Regarding rating scale dimensionality, two aspects were investigated in this study: (1) the extent to which the four criteria within each rating scale worked together to measure one underlying construct and (2) the extent to which the four criteria are distinct from one another to warrant their inclusion. It was evident that the first aspect of dimensionality was well supported by the results from MFRM criterion fit statistics, principal components analysis of residuals, correlations, and principal axis factoring. This further provides support for the common practice of reporting reading-into-writing scores generated from using analytic marking as a single composite score. The second aspect of dimensionality, the distinguishability of the criteria within an analytic rating scale, is also supported by the findings from the MFRM summary statistics and correlations. A few raters, however, perceived overlaps between criteria within a rating scale although these perceived overlaps did not necessarily translate into rating difficulty. Perceived overlaps, or fuzziness, are quite germane to analytic rating scales in general and have been previously investigated (e.g., Brown, 2006; Xi & Mollaun, 2006), and are not an issue specific to the two analytic rating scales in this study. Chan et al. (2015), in their study developing the rating scale used as Rating Scale 1 in this study, also found that some of their raters regarded *Reading for Writing* and *Task Fulfilment* as rather overlapping. This is understandable because the descriptor "achievement of the communicative aim" under the *Task Fulfilment* criterion can be argued to be an overarching goal that inherently takes into account all aspects of performance, including test-takers' ability to use information from the reading texts in their writing, although the *Task Fulfilment* descriptors do not make direct reference to the use of reading material. As for Rating Scale 2, it was not very surprising that raters found the two criteria representing the reading-related construct (i.e., *Source Use* and *Development of Ideas*) often hard to distinguish, because the two make direct reference to how test-takers use information from the source texts.

Raters' perceived overlaps should not invalidate the usability of analytic rating scales for reading-into-writing assessment; rather, this information is invaluable to inform rating scale development and/or revision, rater training, rater monitoring, and score reporting. Especially when test designers decide to operationalize the reading-related aspects into more than one criterion in an analytic rating scale, as in the case of Rating Scale 2, these criteria must be adequately distinguishable, and empirical evidence—both quantitative and qualitative—must be sought to support the decision. However, it is perhaps impossible to completely eliminate perceived overlaps among criteria within an analytic rating scale, as the different criteria are meant to work together to measure a unified construct (Knoch et al., 2020). What is important is to make raters aware that perceived overlaps should not lead to score interdependence and halo effects, and such issues should be clarified during rater training. Another relevant consideration is whether subscores will be reported individually or as a composite score. If individually, then separability is important; if summed, then they need to represent a single measurement dimension.

Conclusion

While analytic rating scales are frequently employed to evaluate integrated tasks, little is known about how specific operationalizations of the integrated language construct in

such rating scales may affect rating quality, and by extension score inferences and uses. To the best of our knowledge, this study is the first to have empirically investigated the potential effect of two different ways of integrated-construct operationalization in analytic rating scales—in this case, in the context of testing reading-into-writing. The two approaches explored concerned scales in which reading and source use aspects of integrated task performances were exclusively captured in one or more distinct criteria versus scales where such aspects were (also) interwoven with other aspects of integrated task performance (e.g., linguistic resources, idea development) within criteria. By means of two scales exemplifying these two approaches and a group of raters' ratings of reading-into-writing performances using each scale, it was found that either scale type generally functioned well and represented one underlying integrated construct. At the same time, it appeared that rating reading-related aspects of reading-into-writing performances remains more challenging for raters than judging writing-related aspects, regardless of the type of analytic rating scale. Also, while raters' scoring data indicate that the individual criteria within either type of scale are distinguishable, raters do not always perceive this to be the case. The two types of scales differed, however, in the extent to which they led raters to uniform interpretations of performance difficulty levels. Specifically, the analytic scale which operationalized the reading and source use aspects in an exclusive manner (distinct from writing-related criteria; Rating Scale 1) resulted in more uniform difficulty level interpretations by raters. This type of scale might therefore be slightly more preferable than those combining multiple aspects within a criterion.

This study is not without limitations. Namely, performance ratings were conducted for one reading-into-writing task only; it remains to be explored how the two types of analytic rating scales function with other versions of the task and, importantly, with other types of reading-into-writing tasks (e.g., different input or output genres, different requirements to draw on the input). In addition, apart from differing in terms of the way the reading-related aspects of the task were operationalized, the two rating scales employed in this study contained some differences in the length of the descriptors in each scale; it is possible that this also contributed to raters' use and perceptions of the two scales.

A meaningful avenue for follow-up research could specifically focus on the evaluation of the reading-related aspects of reading-into-writing performances, to shed more light on the differential difficulty and scale structure of this aspect, as identified in the present study. Such research might benefit from the use of verbal report methodology with raters (e.g., think-alouds or stimulated recalls) and/or eye-tracking technology to record raters' eye movements on the scales and performances, in order to provide further insights into rater cognition and decision-making processes.

Acknowledgements

We would like to thank Trinity College London for giving us access to the reading-into-writing task and examinee scripts used in this study. We would also like to thank the participating raters.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/ or publication of this article: The authors would like to acknowledge the funding assistance of the British Council and Duolingo, which enabled the first author to conduct part of the study under a British Council's Assessment Research Award 2019 and a Duolingo English Test's 2020 Doctoral Award. Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of Trinity College London, the British Council or Duolingo.

ORCID iDs

Santi B. Lestari https://orcid.org/0000-0001-9846-3430
Tineke Brunfaut https://orcid.org/0000-0001-8018-8004

References

Andrich, D. A. (1978). A rating scale formulation for ordered response categories. *Psychometrika*, *43*, 561–573. https://doi.org/10.1007/BF02293814

Brown, A. A. (2006). *An examination of the rating process in the revised IELTS Speaking Test*. https://www.ielts.org/-/media/research-reports/ielts_rr_volume06_report2.ashx

Chan, S. (2013). *Establishing the validity of reading-into-writing tasks for the UK academic context* [Doctoral thesis]. University of Bedforshire. https://uobrep.openrepository.com/handle/10547/312629

Chan, S. (2018). *Defining integrated reading-into-writing constructs: Evidence at the B2-C1 interface*. Cambridge University Press.

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, *26*, 20–37. https://doi.org/10.1016/j.asw.2015.07.004

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*(1), 1–8. https://doi.org/10.1080/15434303.2011.622016

Cumming, A. (2014). Assessing integrated skills. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 216–229). Wiley-Blackwell. https://doi.org/10.1002/9781118411360.wbcla131

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*(1), 5–43. https://doi.org/10.1016/j.asw.2005.02.001

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESF/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*, 67–96. https://doi.org/10.1109/AERO.2009.4839648

Delaney, Y. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes*, *7*(3), 140–150. https://doi.org/10.1016/j.jeap.2008.04.001

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing*, *21*, 56–73. https://doi.org/10.1016/j.asw.2014.03.002

Gebril, A., & Plakans, L. (2016). Source-based tasks in academic writing assessment: Lexical diversity, textual borrowing and proficiency. *Journal of English for Academic Purposes*, *24*, 78–88. https://doi.org/10.1016/j.jeap.2016.10.001

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. *Language Testing*, *28*(2), 179–200. https://doi.org/10.1177/0265532210384252

Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, option and directions*. Equinox Publishing.

Knoch, U., Ying, B., Elder, C., Flynn, E., Huisman, A., Woodward-kron, R., Manias, E., & McNamara, T. (2020). "I will go to my grave fighting for grammar": Exploring the ability of language-trained raters to implement a professionally-relevant rating scale for writing. *Assessing Writing*, *46*, Article 100488. https://doi.org/10.1016/j.asw.2020.100488

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), Article 878. https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2015a). *Facets* (Version 3.71.4) [Computer software]. https://www.winsteps.com/facets.htm

Linacre, J. M. (2015b). *Winsteps* [Computer software]. https://www.winsteps.com/winsteps.htm

Linacre, J. M. (2020). *Facets program manual 3.83.4*. https://www.winsteps.com/a/Facets-Manual.pdf

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment*. Oxford University Press.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

Ohta, R., Plakans, L. M., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing*, *38*, 21–36. https://doi.org/10.1016/j.asw.2018.08.001

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, *13*(2), 111–129. https://doi.org/10.1016/j.asw.2008.07.001

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561–587. https://doi.org/10.1177/0265532209340192

Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing*, *17*(1), 18–34. https://doi.org/10.1016/j.asw.2011.09.002

Plakans, L., & Gebril, A. (2015). *Assessment myths: Applying second language research to classroom teaching*. University of Michigan Press.

Shin, S. Y., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores? *Language Testing*, *32*(2), 259–281. https://doi.org/10.1177/0265532214560257

Spivey, N. N. (1997). *The Constructivist metaphor: Reading, writing and the making of meaning*. Brill.

Staples, S., Biber, D., & Reppen, R. (2018). Using corpus-based register analysis to explore the authenticity of high-stakes language exams: A register comparison of TOEFL iBT and disciplinary writing tasks. *Modern Language Journal*, *102*(2), 310–332. https://doi.org/10.1111/modl.12465

Trinity College London (2015). Integrated Skills in English (ISE): Specifications – reading & writing, ISE Foundation to ISE III. Trinity College London. https://www.trinitycollege.com/resource/?id=6299

Trinity College London (n.d.). Integrated Skills in English ISE II: Reading & writing exam, sample paper 3. Trinity College London. https://www.trinitycollege.com/resource/?id=7202

Wang, J., Engelhard, G., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing*, *33*, 36–47. https://doi.org/10.1016/j.asw.2017.03.003

Wang, P. (2018). *Investigating test-takers' cognitive processes while completing an integrated reading-to-write task: Evidence from eye-tracking, stimulated recall, and questionnaire* [Unpublished doctoral thesis]. Lancaster University.

Wind, S. A. (2020). Do raters use rating scale categories consistently across analytic rubric domains in writing assessment? *Assessing Writing*, *43*, Article 100416. https://doi.org/10.1016/j.asw.2019.100416

Wind, S. A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing*, *18*(4), 278–299. https://doi.org/10.1016/j.asw.2013.09.002

Wind, S. A., & Jones, E. (2019). The effects of incomplete rating designs in combination with rater effects. *Journal of Educational Measurement*, *56*(1), 76–100. https://doi.org/10.1111/jedm.12201

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Xi, X., & Mollaun, P. (2006). *Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST)*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/RR-06-07.pdf

Yu, G. (2013). From integrative to integrated language assessment: Are we there yet? *Language Assessment Quarterly*, *10*(1), 110–114. https://doi.org/10.1080/15434303.2013.766744